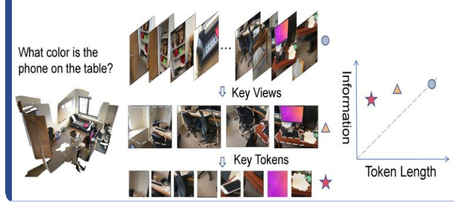
KeyVT: Zero-Shot 3D Question Answering via Hierarchical View to Token Transportation

Dongsheng Wang · Dawei Su · Hui Huang*
College of Computer Science and Software Engineering · Shenzhen University, China

ICML International Conference On Machine Learning
2026 · South Korea

Paper Code

Motivation

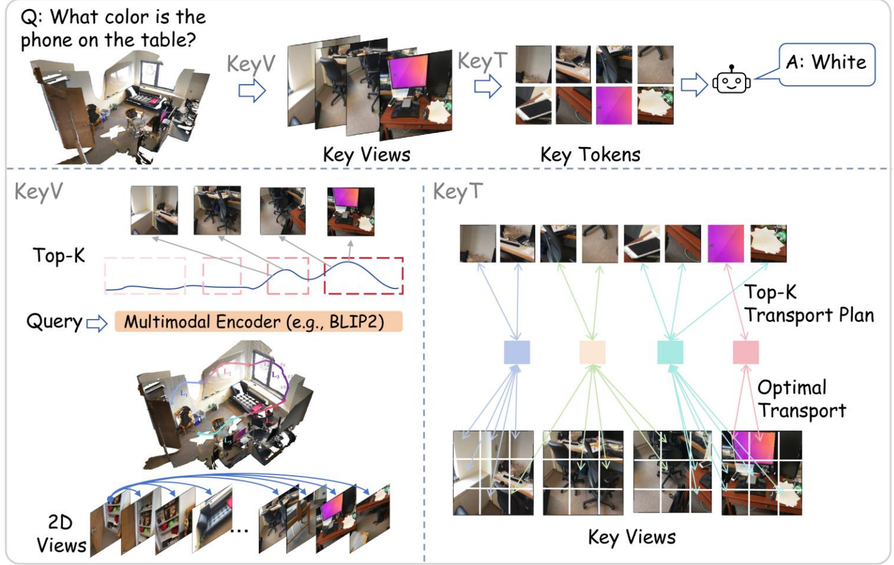


Problem: Token limits in 3D-QA cause spatial context loss and severe visual redundancy.

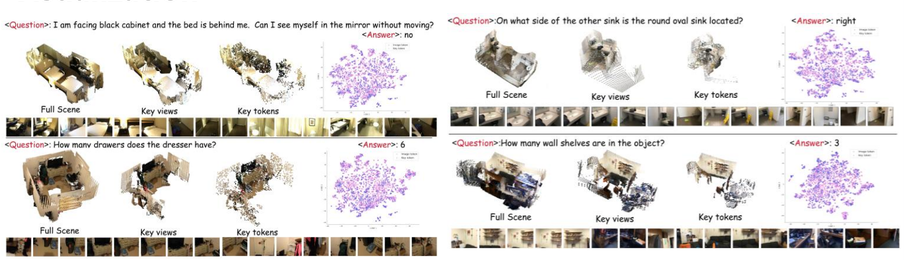
Method: KeyVT selects spatially consistent views and compresses redundant tokens via optimal transport.

Impact: Maximizes diverse 3D context and frees input capacity without increasing computational overhead.

Framework of KeyVT



Visualization



Method

KeyV:Geometry-aware Key View Selection

View Distance Calculates spatial trajectories by combining camera position offsets and rotational angular differences to ensure spatial consistency.

$$D(V_0, V_i) = ||\mathbf{C}_0 - \mathbf{C}_i|| + \theta(\mathbf{R}_0, \mathbf{R}_i),$$

Sub-scene Relevance Score Evaluates sub-scene semantic importance by aggregating the maximum and average vision-language similarity scores.

$$r_l = \max(O_l) + \text{mean}(O_l), \quad O_{l,i\in L_l} = \text{BLIP2}(V_i, Q),$$

Adaptive View Allocation Dynamically assigns the optimal number of key views to different sub-scenes based on their size and semantic relevance.

$$N_l = \lfloor K \cdot \frac{W_l}{\sum_{l'}^{|L|} W_{l'}} \rfloor, \quad W_l = r_l \cdot \sqrt{|L_l|},$$

Key Insight: Instead of relying on isolated semantic retrieval, incorporating camera parameters preserves crucial spatial coherence for 3D reasoning.

KeyT:Optimal-Transport based Key Tokens compression

Optimal Transport (OT) Distance Formulates token compression by minimizing the transportation cost between redundant view tokens and compact virtual tokens.

$$\mathbf{P} = \sum_{n=1}^{N} \alpha_n \boldsymbol{e}_n, \quad \mathbf{Q} = \sum_{m=1}^{M} \beta_m \boldsymbol{c}_m, \qquad d_{\mathbf{C}}(\mathbf{P}, \mathbf{Q}) = \min_{\mathbf{T}\in U(\alpha,\beta)} < \mathbf{T}, \mathbf{C} >,$$

Sinkhorn Distance Introduces entropy regularization to the optimal transport problem, enabling fast and fully differentiable gradient-based token updates.

$$d_{\mathbf{C},\gamma}(\mathbf{P}, \mathbf{Q}) = \min_{\mathbf{T}\in U_\gamma(\alpha,\beta)} < \mathbf{T}, \mathbf{C} >,$$

Real Token Grounding Bridges the embedding gap by selecting the actual image patch tokens that best correspond to the optimized virtual tokens.

$$\text{Nei}(\boldsymbol{c}_m, \mathbf{T}) = \{\boldsymbol{e}_i | i \in \text{Top-K}(\frac{S}{M}, \mathbf{T}_{\cdot,m})\},$$

Key Insight: While clustering-based models are sensitive to densely populated features, learning tokens under the optimal transport (OT) theory leverages its ability to model geometric structures, enabling the model to discover diverse and representative tokens.

Experiments

Results on ScanQA & SQA3D

Methods	ScanQA (val)					SQA3D (test)	input frames
	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	EM-1	
ScanQA	30.2	10.1	13.1	33.3	64.9	47.2	–
SQA3D	30.5	11.2	13.5	34.5	-	46.6	–
3D-Vista	-	-	13.9	35.7	-	48.5	–
3D-LLM	39.3	12.0	14.5	35.7	69.4	-	–
LL3DA	-	13.5	15.9	37.3	76.8	-	–
Chat-Scene	43.2	14.3	18.0	41.6	87.7	54.6	–
3D-LLaVA	-	17.1	18.4	43.1	92.6	54.5	–
cdViews (LLaVA-OV-7B)	42.6	-	-	46.8	94.0	56.9	9
Video-3D (LLavAVideo-7B)	-	-	-	-	96.4	57.0	8
Video-3D (LLavAVideo-7B)	-	-	-	-	100.6	57.8	16
Video-3D (LLavAVideo-7B)	47.1	16.2	19.8	49.0	102.1	58.6	32
LLaVA-OV-7B	38.5	10.3	16.4	43.1	84.0	53.9	8
+AKS	41.5	**13.6**	17.5	45.2	90.2	56.2	8
+DivPrune	40.3	13.1	17.3	45.1	89.5	53.8	8^{eq}
+FLoC	41.3	13.7	17.6	45.6	91.1	54.8	8^{eq}
+KeyVT (Ours)	**41.7**	12.4	**17.9**	**46.7**	**93.8**	**57.1**	8^{eq}
LLaVAVideo-7B	43.9	12.2	18.4	45.9	93.4	54.7	8
+AKS	45.7	12.6	19.2	48.1	99.0	57.3	8
+DivPrune	46.2	12.3	19.3	48.5	99.1	55.8	8^{eq}
+FLoC	46.4	10.9	19.3	48.5	99.4	57.2	8^{eq}
+KeyVT (Ours)	**46.6**	**14.5**	**19.4**	**48.8**	**100.7**	**57.9**	8^{eq}
Qwen2.5VL-7B	26.6	7.5	12.4	34.2	63.4	46.0	8
+AKS	28.3	7.5	13.2	36.4	67.2	49.5	8
+DivPrune	27.9	8.3	13.2	36.5	67.3	50.1	8^{eq}
+FLoC	28.4	**8.6**	13.5	37.0	68.5	50.2	8^{eq}
+KeyVT (Ours)	**29.7**	8.5	**13.5**	**37.3**	**69.6**	**50.4**	8^{eq}

Results on VSI-Bench

Methods	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
		Numerical Answer				Multiple-Choice Answer			
Proprietary Models (API)									
Gemini-1.5 Flash	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
GPT-4o	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Open-source Models									
LLaVA-OneVision-7B	31.4	34.7	20.6	47.3	18.1	40.4	32.4	**32.5**	24.9
+ AKS	32.8	43.1	19.8	46.5	24.2	**41.3**	36.8	29.1	21.6
+ FLoC	33.1	**47.1**	**21.2**	48.0	26.5	37.9	36.1	28.9	19.1
+ DivPrune	33.4	43.2	21.1	**49.4**	**28.6**	37.9	37.4	28.5	21.0
+See&Trek	33.0	32.0	17.0	39.8	27.8	39.0	**40.6**	31.9	**35.7**
+KeyVT (Ours)	**33.9**	46.6	20.6	48.8	26.2	38.7	38.8	29.4	21.7
LLaVAVideo-7B	32.2	40.0	15.1	46.8	25.2	38.9	39.4	25.7	26.5
+ AKS	33.2	41.8	14.6	48.6	**26.8**	**41.0**	**41.3**	33.0	18.8
+ FLoC	34.0	44.1	22.2	48.7	25.7	39.4	39.2	32.5	20.3
+ DivPrune	32.8	39.8	15.2	**49.1**	25.5	39.9	40.0	30.9	**22.0**
+KeyVT (Ours)	**34.6**	**44.5**	**23.8**	47.1	23.6	40.1	40.9	**35.6**	21.4
Qwen2.5-VL-7B	27.3	13.0	14.4	35.9	21.3	**36.9**	37.9	29.9	29.6
+ AKS	36.2	**48.0**	23.5	**47.0**	28.9	34.1	37.6	**38.1**	32.4
+ FLoC	36.6	46.7	24.8	45.3	33.2	36.3	38.5	35.1	32.1
+ DivPrune	36.1	46.3	25.3	45.1	31.1	33.7	38.6	36.1	33.1
+See&Trek	29.0	13.6	14.7	35.4	23.6	33.4	**41.3**	30.4	39.2
+KeyVT (Ours)	**37.0**	41.4	**26.7**	46.0	**34.8**	31.7	38.7	35.6	**40.8**